AN OPTIMAL PROPERTY OF PRINCIPAL COMPONENTS

BY J. N. DARROCH

University of Michigan

1. Introduction and summary. Let $x' = (x_1, x_2, \cdots, x_p)$ be a random vector and let $E[x] = 0$, $E[xx'] = \Sigma = (\sigma_{ij})$ where we assume that Σ is non-singular. Further let

$$\Sigma = T\Lambda T' = (t_1, t_2, \cdots, t_p)\begin{bmatrix} \lambda_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \lambda_2 & \cdot & \cdot & \cdot & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdot & \cdot & \cdot & \lambda_p \end{bmatrix}\begin{bmatrix} t'_1 \\ t'_2 \\ \vdots \\ t'_p \end{bmatrix}$$

where $TT' = I$ and where we suppose that the eigenvalues of Σ are in order of decreasing magnitude, that is $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p > 0$. The principal components of x, namely $u_1 = t_1'x$, $u_2 = t_2'x$, $\cdots$, $u_p = t_p'x$ were introduced by Hotelling (1933) who characterised them by certain optimal properties. Since then Girshick (1936), Anderson (1958) and Kullback (1959) have characterised the principal components by slightly different sets of optimal properties. Thus Anderson shows that u_1 is the linear function $\alpha_1'x$ having maximum variance subject to $\alpha_1'\alpha_1 = 1$; u_2 is the linear function $\alpha_2'x$ which is uncorrelated with u_1 and has maximum variance subject to $\alpha_2'\alpha_2 = 1$; and so on.

The above mentioned characterisations have two properties in common; they introduce the principal components one by one and, more importantly, the optimal properties hold only with-in the class of linear functions of $x_1, x_2, \cdots, x_p$.

In the following theorem the first k principal components are characterized by an optimal property within the class of *all* random variables.

2. The optimal property. Before stating the main result we note that, since

$$TT' = t_1t_1' + t_2t_2' + \cdots + t_pt_p' = I,$$

therefore

$$x = t_1u_1 + t_2u_2 + \cdots + t_pu_p.$$

THEOREM. *Let A be any $p \times k$ matrix and let $f' = (f_1, f_2, \cdots, f_k)$ be any random vector. Then*

$$\mathfrak{J}_1 = \text{trace } E[(x - Af)(x - Af)'] = E[\textstyle\sum_{i=1}^{p} (x_i - (Af)_i)^2]$$

is minimized with respect to A and f when and only when

$$Af = t_1u_1 + t_2u_2 + \cdots + t_ku_k,$$

and the minimum value of $\mathfrak{J}_1$ is $\sum_{i=k+1}^{p} \lambda_i$.

Received 24 February 1965.

PROOF. Without loss of generality suppose that $E[ff'] = I$ and let $E[xf'] = B$. Then B must satisfy the condition

$$\begin{bmatrix} \Sigma & B \\ B' & I \end{bmatrix} \text{ is non-negative definite,} \tag{1}$$

since the matrix in (1) is the covariance matrix of $(x', f') = (x_1, x_2, \cdots, x_p, f_1, f_2, \cdots, f_k)$. Now, if Γ_{22} is positive definite, a necessary and sufficient condition for

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}$$

to be non-negative definite is that $\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$ is non-negative definite. Therefore condition (1) is equivalent to

$$\Sigma - BB' \text{ is non-negative definite.} \tag{2}$$

Now

$$E[x - Af)(x - Af)']$$
$$= \Sigma - AB' - BA' + AA' = \Sigma - BB' + (A - B)(A - B)'.$$

Since tr $(A - B)(A - B)' \geqq 0$ with equality if and only if $A = B$, it follows that the minimum of $\mathfrak{J}_1$ with respect to A is $\mathfrak{J}_2 = \operatorname{tr}(\Sigma - BB')$. As in Section 1 write $\Sigma = T\Lambda T'$, and define

$$C = \Lambda^{-\frac{1}{2}}T'B. \tag{3}$$

Then

$$\begin{aligned} \mathfrak{J}_2 &= \operatorname{tr}(T\Lambda T' - T\Lambda^{\frac{1}{2}}CC'\Lambda^{\frac{1}{2}}T') \\ &= \operatorname{tr}(T'T(\Lambda - \Lambda^{\frac{1}{2}}CC'\Lambda^{\frac{1}{2}})) \\ &= \operatorname{tr}(\Lambda - \Lambda^{\frac{1}{2}}CC'\Lambda^{\frac{1}{2}}) \\ &= \operatorname{tr}\Lambda(I - CC'). \end{aligned}$$

Since C is a $p \times k$ matrix, CC' is at most of rank k. Therefore we can find P such that $CC' = PDP'$, $PP' = I$, where D is a diagonal matrix of the form

$$D = \begin{bmatrix} d_1 & & & & & \\ & \ddots & & & 0 & \\ & & d_k & & & \\ & 0 & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

We can now write

$$\begin{aligned}\mathfrak{J}_2 &= \operatorname{tr} \Lambda\,(I - PDP') \\ &= \textstyle\sum_{i=1}^{p} \lambda_i - \sum_{i=1}^{p} \lambda_i \sum_{j=1}^{k} p_{ij}^2\, dj\,.\end{aligned}$$

Condition (2) has meanwhile become

$$I - D \text{ is non-negative definite.} \tag{4}$$

Therefore $\mathfrak{J}_2$ is minimised subject to (4) by choosing $d_1 = \cdots = d_k = 1$. It remains to minimise

$$\mathfrak{J}_3 = \textstyle\sum_{i=1}^{p} \lambda_i - \sum_{i=1}^{p} \lambda_i w_i$$

with respect to $w_1, w_2, \cdots, w_p$ where

$$w_i = \textstyle\sum_{j=1}^{k} p_{ij}^2\,.$$

Because P is orthogonal,

$$0 \leqq w_i \leqq 1, \qquad \textstyle\sum_{i=1}^{p} w_i = \sum_{i=1}^{p} \sum_{j=1}^{k} p_{ij}^2 = k. \tag{5}$$

Therefore $\mathfrak{J}_3$ is minimised with respect to (5) by choosing

$$w_1 = \cdots = w_k = 1, \qquad w_{k+1} = \cdots = w_p = 0 \tag{6}$$

and the minimum value of $\mathfrak{J}_3$ is therefore $\sum_{i=k+1}^{p} \lambda_i$. Equations (6) are equivalent to

$$P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}$$

where P_{11} is $k \times k$.

Retracing our steps we find that $\mathfrak{J}_2$ is minimised when

$$C = \begin{bmatrix} Q \\ 0 \end{bmatrix}$$

where Q is any orthogonal $k \times k$ matrix. Therefore $\mathfrak{J}_2$ is minimised when

$$B = T\Lambda^{\frac{1}{2}} \begin{bmatrix} Q \\ 0 \end{bmatrix} = G \text{ say.} \tag{7}$$

So far we have shown that, if g is a random vector with the properties that $E[xg'] = G$, $E[gg'] = I$, then the minimum value of $\mathfrak{J}_1$, namely $\sum_{i=k+1}^{p} \lambda_i$, is attained by taking Af equal to Gg.

Now let $v = Hx$ where H is a $k \times p$ matrix defined by $E[xv'] = G$. Thus

$$H = G'\Sigma^{-1} \tag{8}$$

and we see that

$$\begin{aligned} E[gg'] &= H\Sigma H' = G'\Sigma^{-1}\Sigma\Sigma^{-1}G \\ &= (Q'0)\Lambda^{\frac{1}{2}}T'\Sigma^{-1}T\Lambda^{\frac{1}{2}}\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ &= (Q'0)\Lambda^{\frac{1}{2}}\Lambda^{-1}\Lambda^{\frac{1}{2}}\begin{bmatrix} Q \\ 0 \end{bmatrix} \\ &= I. \end{aligned}$$

Thus v satisfies all the conditions on g. Moreover, neglecting differences which have zero probability measure, it is the *only* vector to do so.

For

$$\begin{aligned} E[(v - g)(v - g)'] &= E[(Hx - g)(Hx - g)'] \\ &= H\Sigma H' - HG - G'H' + I \\ &= 0 \end{aligned}$$

by (7) and (8).

Finally, it is easily verified that

$$\begin{aligned} v &= T\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} T'x \\ &= t_1(t_1' x) + \cdots + t_k(t_k' x). \end{aligned}$$

Thus $\mathfrak{J}_1$ is minimised uniquely by taking

$$Af = t_1u_1 + \cdots + t_ku_k .$$

ACKNOWLEDGMENT. I wish to thank Professor R. Berk for some useful discussion concerning the proof of this theorem.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

GIRSHICK, M. A. (1936). Principal components. *J. Amer. Statist. Assoc.* **31** 519–528.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24** 417–444; 498–520.

KULLBACK, S. (1959). *Information Theory and Statistics*, Wiley, New York.